Delineating environmental envelopes to improve mapping of species distributions, via a hurdle model with CART &/or MaxEnt

R. Pirathiban [a], **K. J. Williams** [b], **S. J. Low Choy** [ac]

[a]*School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD 4001, Australia*
[b]*CSIRO Land and Water, Canberra, ACT 2601, Australia*
[c]*Griffith Social and Behavioural Research College, Griffith University, Mt Gravatt, QLD 4122, Australia*
Email: r.jeyapalan@qut.edu.au

Abstract: Species distribution modelling (SDM) typically analyses species' presence together with some form of absence information. Ideally absences comprise observations or are inferred from comprehensive sampling. When such information is not available, then pseudo-absences are often generated from the background locations within the study region of interest containing the presences, or else absence is implied through the comparison of presences to the whole study region, *e.g.* as is the case in Maximum Entropy (MaxEnt) or Poisson point process modelling.

However, the choice of *which* absence information to include can be both challenging and highly influential on SDM predictions (e.g. Oksanen and Minchin, 2002). In practice, the use of pseudo- or implied absences often leads to an imbalance where absences far outnumber presences. This leaves analysis highly susceptible to 'naughty-noughts': absences that occur beyond the envelope of the species, which can exert strong influence on the model and its predictions (Austin and Meyers, 1996). Also known as 'excess zeros', naughty noughts can be estimated via an overall proportion in simple hurdle or mixture models (Martin et al., 2005). However, absences, especially those that occur beyond the species envelope, can often be more diverse than presences. Here we consider an extension to excess zero models. The two-staged approach first exploits the compartmentalisation provided by classification trees (CTs) (as in O'Leary, 2008) to identify multiple sources of naughty noughts and simultaneously delineate several species envelopes. Then SDMs can be fit separately within each envelope, and for this stage, we examine both CTs (as in Falk et al., 2014) and the popular MaxEnt (Elith et al., 2006).

We introduce a wider range of model performance measures to improve treatment of naughty noughts in SDM. We retain an overall measure of model performance, the area under the curve (AUC) of the Receiver-Operating Curve (ROC), but focus on its constituent measures of false negative rate (FNR) and false positive rate (FPR), and how these relate to the threshold in the predicted probability of presence that delimits predicted presence from absence. We also propose error rates more relevant to users of predictions: false omission rate (FOR), the chance that a predicted absence corresponds to (and hence wastes) an observed presence, and the false discovery rate (FDR), reflecting those predicted (or potential) presences that correspond to absence. A high FDR may be desirable since it could help target future search efforts, whereas zero or low FOR is desirable since it indicates none of the (often valuable) presences have been ignored in the SDM.

For illustration, we chose *Bradypus variegatus*, a species that has previously been published as an exemplar species for MaxEnt, proposed by Phillips et al. (2006). We used CTs to increasingly refine the species envelope, starting with the whole study region (E0), eliminating more and more potential naughty noughts (E1–E3). When combined with an SDM fit within the species envelope, the best CT SDM had similar AUC and FPR to the best MaxEnt SDM, but otherwise performed better. The FNR and FOR were greatly reduced, suggesting that CTs handle absences better. Interestingly, MaxEnt predictions showed low discriminatory performance, with the most common predicted probability of presence being in the same range (0.00-0.20) for both true absences and presences. In summary, this example shows that SDMs can be improved by introducing an initial hurdle to identify naughty noughts and partition the envelope before applying SDMs. This improvement was barely detectable via AUC and FPR yet visible in FOR, FNR, and the comparison of predicted probability of presence distribution for pres/absence.

Keywords: *Naughty-noughts, excess zeros, species distribution modelling, predictive performance, misclassification rate*

1 INTRODUCTION

Evaluation of the underlying statistical methods used for species distribution modelling (SDM) has become increasingly addressed in landscape ecology (Elith et al., 2006; Guisan et al., 2013), particularly when predicting beyond the range of the input data, for example under climate change (Guisan et al., 2006; Elith et al., 2010). For SDMs the extent to which the modelled distributions reflect real-world species distributions depends on the extent to which the input data provides a complete yet representative coverage of those factors affecting the geographic and environmental distribution of the species. Indeed, a conservative approach would include all areas (and habitats) where the target species might be found. This strategy can improve completeness, but at great cost to representativeness, potentially leading to an excess of absences. Such mis-representation of absence may lead to substantial bias in parameter estimates as well as predictions (Oksanen and Minchin, 2002).

This mis-representation often occurs when too many areas are included where the species cannot exist, together with a large number of real or pseudo-absence records. This leads to many 'ambiguous' absences (Low Choy, 2001) and makes it challenging to differentiate between 'structural' and 'random' absences (Martin et al., 2005). Informally called 'naughty-noughts' (Austin and Meyers, 1996), these structural absences occur beyond the feasible range of the species whereas 'random' absences can occur when the species inhabits only some portion of the suitable habitat, when the species is not always detected when present, or the sampling strategy was otherwise limited.

Over the last two decades, some methodologies have been developed to account for excess zeros, and include both sampling and modelling approaches. A common sampling approach defines a geographical partition (Osborne and Suárez Seoane, 2002) of the data by drawing concentric rings around the centroid of the species distribution, thereby omitting areas that are clearly unsuitable that contain many 'naughty-naughts'. A more sophisticated environmental partition buffers the geographic range for the species capturing all actual presences (Williams, 1998). Environmental thresholds can also be determined to delimit the environmental factors corresponding to uninhabited areas, and hence avoid excessive 'naughty-noughts' (Austin and Meyers, 1996). Phillips et al. (2009) investigate the idea of 'target group absences' through the use of all occurrences within a target group of several species as pseudo-absence data. Although these approaches solve some of the 'naughty-naughts' issues, the consequences of limiting the data without undertaking any formal procedures are so severe that they can produce ecologically non-sensible species response curves (Oksanen and Minchin, 2002).

The alternative is to use a modelling approach, such as zero inflation modelling, to focus on identifying the likely 'naughty-noughts', using simple mixture or hurdle models (Martin et al., 2005). However these simple models treat all 'naughty-noughts' in the same way. To allow different sources of absences, the hurdle model has been extended to two-steps (O'Leary, 2008; Falk et al., 2014). In the first step, environmental envelopes are defined using CART with unequal penalties on misclassification errors via recursive partitioning (O'Leary, 2008) or a Bayesian algorithm (Falk et al., 2014). In the second step, a GLM or GAM regression model can be used to fit an SDM within each environmental envelope. These models allow ecologists to examine the underlying reasons for 'naughty-noughts', and can lead to large gains in predictive performance.

This study further extends the hurdle model by using the popular MaxEnt (Phillips et al., 2006) as the SDM. To illustrate the approach, we consider a case study, which involves *Bradypus variegatus* (Phillips et al., 2006; Hijmans and Elith, 2015), a species of sloth with very few presence records and a random selection of pseudo-absences across South America.

2 METHOD

2.1 Classification and regression trees

Classification and regression trees (CART) are constructed by repeatedly defining branches to split the data into two groups, with each branch defined by a splitting rule based on a single explanatory variable. At each split the data is partitioned into two mutually exclusive subsets, each more homogeneous than before the split. The splitting procedure is then recursively applied to each subset separately. The objective is to partition the response into homogeneous subsets, but also to keep the tree reasonably small. At each step in fitting a tree, selection of the splitting rule is determined by optimising some measure of homogeneity (e.g. Gini or Information index).

In this study, the focus is on classification trees (CT) since the response variable of the considered case study is binary (presence and absence). Trees are popular due to their intuitive visual representation: all presences

and absences are filtered from the root node, which represents the undivided data, at the top, along a sequence of branches before reaching a terminal leaf node, which here defines the environmental profile of presences or absences.

2.2 Identifying and eliminating naughty noughts to delineate environmental envelopes

Environmental profiles of presences and absences are defined with the aim of identifying uninhabitable regions and hence minimize the chance that presences are predicted as absences. This can be achieved by fitting a classification tree using the recursive partitioning algorithm in R (Therneau et al., 1997), which allows us to assign greater penalty to misclassification of true presences as absences compared to misclassification of true absences as presences. Thus one or more environmental profiles (nodes) have zero or near zero false omission rate (FOR), being the chance that predicted absences correspond to true presences. This contrasts with the approach outlined in O'Leary (2008) which focussed on controlling false negative rate (FNR), being the chance that a true presence is predicted to be present.

Other profiles (nodes) that contain presences mixed with absences can be interpreted as environmental envelopes. Following O'Leary (2008) we gradually refine the environmental envelopes, by progressively eliminating more and more uninhabitable landscapes.

2.3 MaxEnt and CT for fitting SDMs within envelopes

At the next step, an SDM is fit within each envelope. Here we examine two different SDMs — the popular MaxEnt (Elith et al., 2006; Phillips et al., 2006), and the simpler classification tree model (O'Leary, 2008; Falk et al., 2014).

MaxEnt, applies the maximum entropy principle to describe the relative likelihood of each environmental predictor independently to define the features of sites at which the species occurs compared to the features of the environment as a whole. This contrasts with GLMs and CTs which model the relative likelihood of presences given the environmental characteristics. Here, MaxEnt is fitted within each environmental envelope using the R package `dismo` (Hijmans and Elith, 2015) under default setting. After eliminating the potential 'naughty-naughts', for each envelope, we provided the remaining sites with absence records as background sample (which are a random or regular sample of the landscape with no presences). As MaxEnt's raw output represents the relative suitability, it needs to be post-processed to reverse the logic to estimate occupancy probabilities, e.g. via estimation of prevalence (Guillera-Arroita et al., 2014).

A CT is also fitted within each environmental envelope, as well as the whole region, to identify the habitable and inhabitable environmental profiles. We ensured that the most likely prediction for observed presences was presence by placing a higher cost on the misclassification of presences. This aim is different compared to isolation of 'naughty-noughts'.

2.4 Evaluation measures

Predictions were obtained by applying CT or MaxEnt SDMs within the series of three environmental envelopes, and then combined with the predictions of the eliminated 'naughty-naughts'. These were also compared to the more usual application of CT or MaxEnt to the entire dataset. This provided a set of eight models to evaluate the model performance: four environmental envelopes (gradually eliminating more and more 'naughty-noughts' from the overall dataset) and two SDMs (CT, MaxEnt).

The discrimination capacity of the models is examined through three diagnostic plots: receiver operating characteristic (ROC) curve (Pearce and Ferrier, 2000; Metz, 1978); the diagnostic error trade-off plot (Fielding and Bell, 1997); and the frequency distribution of the predicted probability of presence for true presences and absences (Pearce and Ferrier, 2000; Murphy and Winkler, 1977). In addition we consider four misclassification errors, all based on the confusion (or error) matrix (Fielding and Bell, 1997). We separate the overall misclassification rate (MCR) into the usual diagnostic errors: false negative rate (FNR) and false positive rate (FPR), thus ensuring a balance between sensitivity and specificity. In addition we also consider two errors that are largely ignored in the ecological literature: false omission rate (FOR) and false discovery rate (FDR) which can be calculated via the positive predicted value (PPV) and negative predicted value (NPV).

The decision threshold which splits the predicted probabilities into predicted presence and absence for the confusion matrix is determined by selecting the point on the error trade-off plots at which the total misclassification rate is minimised (where FNR and FPR are equal). We have also computed the area under the ROC curve (AUC) which is a model performance measure that is independent of the choice of threshold. In addition

we map the predicted probability of presence across study area.

3 CASE STUDY

3.1 *Bradypus variegatus*

The case study on *Bradypus variegatus* is based on 116 occurrence localities used by Phillips et al. (2006) together with 5000 pseudo-absence points randomly generated from the background. In this paper, we focus on the impact of 'naughty-noughts' and SDM choice on the predictive performance of models. Interpretation of ecological gradients is explored further in Pirathiban et al. (2015).

The environmental covariates used are detailed in Phillips et al. (2006), and can be grouped into three categories: climate, elevation and potential vegetation. The annual and monthly climate variables are annual cloud cover (cld), annual diurnal temperature range (dtr), annual frost frequency (frs), annual vapour pressure (vap), January (pre1), April (pre4), July (pre7), October (pre10) and annual precipitation (pre) and minimum (tmn), maximum (tmx) and mean (tmp) annual temperature. Two other variables are also used in addition to the climatic data: elevation (h_dem) and major habitat types (ecoreg) found in Latin America and the Caribbean.

4 RESULTS

4.1 Naughty noughts

Figure 1 depicts the selected classification tree model fitted for the study region to identify the landscapes which are purely uninhabitable for the species. Misclassification of presences is penalized 110 times more than absences to minimize the FOR. This tree identifies nine uninhabitable landscapes, each containing purely observed absences. These potential 'naughty-noughts' fall into three major groups depending on the ecoregion and the weather during July. For instance, there is a node predicted as absence (A1) containing 1906 observed absences, which identifies ecoregions uninhabited by the species to be: temperate broad leaf and mixed forests, temperate and tropical/subtropical grassland savannas and shrub lands, flooded and montane grasslands, mediterranean scrub, snow/ice/glaciers/rock category, tundra and water (ecoreg = a, c, e, g, i, k, l, m, n and o).

However, the other two major groups of absences fall in these ecoregions depending on the July precipitation, that is above or below 14.5 mm. In drier areas (pre7 $<$ 14.5 mm) there are only two uninhabitable sites whereas the remaining six sites occur in wet areas. The pure absences in group B1 are the second most numerous in terms of pure absences (974) which suggests that *Bradypus variegatus* doesn't prefer ecoregions that are much higher in elevations (hdem $>$ 121.5 m) with an annual diurnal temperature range below 152.5K and even more drier weather in July (pre7 $<$ 12.5 mm). Similarly, several other nodes define uninhabited environmental profiles: B2 (150 absences), C1 (122 absences), C2 (277 absences), C3 (221 absences), C4 (82 absences) and C5 (60 absences) and C6 (68 absences).

4.2 Environmental envelopes

The environmental envelopes (Figure 2) E1, E2 and E3 are determined based on the elimination of the uninhabitable landscapes identified from the classification tree (Figure 1). Environmental envelope E1 is constructed by removing the two pure absences nodes A1 and C1 that are highly distinct from, since not clustered with, the group of leaf nodes that are predicted as presence. It contains all 116 presence observations and 2972 absences. Envelope E2 is the same as E1, but excludes all the pure absence nodes that are not in the same branch as the presence nodes. Thus this envelope removes the absence nodes A1, C1, C3 and C4 and removes an extra 303 absences compared to E1. Finally, E3 comprises solely those nodes that are predicted presences, eliminating all the uninhabitable landscapes, so that this envelope contains all 116 presence observations and only 1140 absences. All absences in envelope E3 occur in the same area as the majority of the presences. However, there is one presence site further inland for which the nearest absence locations have been removed. The majority of the absences in envelope E2 and E1 have a similar range of presence, but some absences occur further inland and E2 contains more absences in the east.

4.3 Fitting MaxEnt and CT within environmental envelopes

The predictions, from both MaxEnt and the CT models, fitted within the environmental envelopes combined with the uninhabitable landscapes are compared against the model fitted to the whole study region. We use discrimination performance measures, model performance measures, diagnostic plots and maps of predictions.

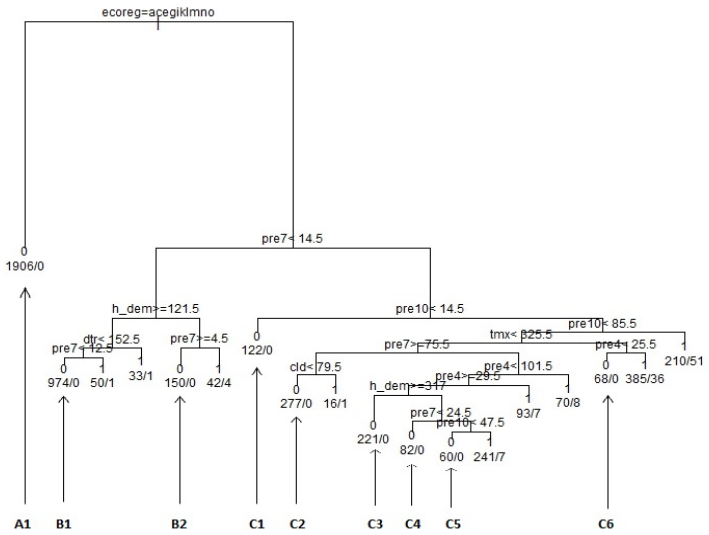


Figure 1: Classification tree that identifies the likely 'naughty-naughts' and the environmental envelopes. Sites are split into two groups based on a splitting rule (such as ecoreg = a, c, e, g, i, k, l, m, n and o in with left = true and right = false), and then each half is split into two sub-groups, and so on. At each leaf node the predicted class of absence or presence is defined by 0 or 1 respectively, with a/p (e.g: 1906/0), the observed number of absences 'a' and presences 'p'. See Section 3.1 for variable definitions.

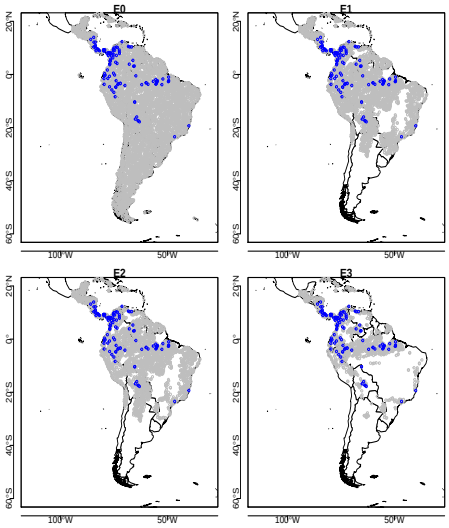
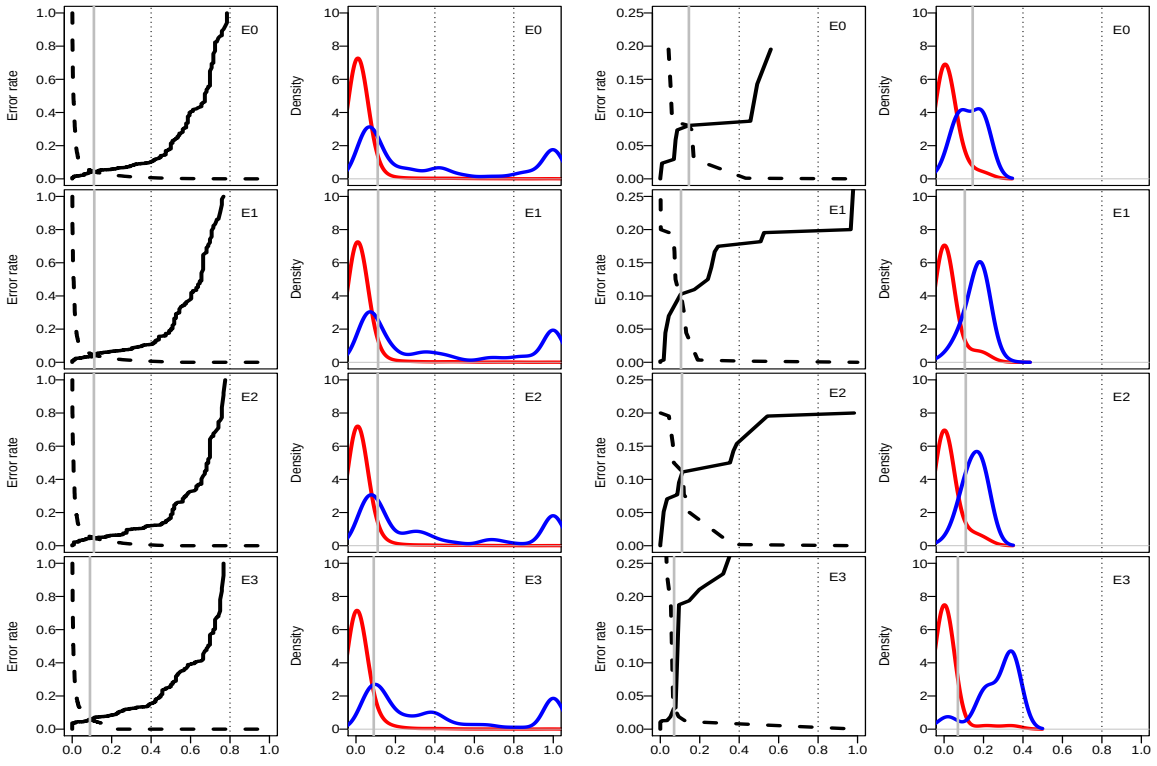


Figure 2: Spatial locations (latitudes and longitudes) of presences (blue) / pseudo-absences (grey) of the species in (E0) study region: 116/5000, (E1) environmental envelope E1: 116/2972, (E2) environmental envelope E2:116/2669 and (E3) environmental envelope E3: 116/1140



(a) MaxEnt **(b)** Classification tree

Figure 3: Error trade off plot - FNR (solid line) verses FPR (dashed line) and the density curves of the predicted probability of presences (blue) and absences (red)

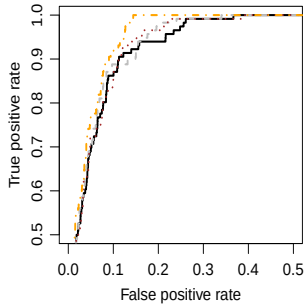
Table 1: Diagnostic performance measures

(a) MaxEnt

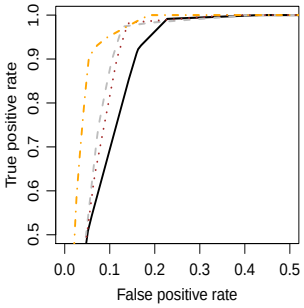
Dataset	AUC	FNR	FPR	FOR	FDR
E0	0.953	0.413	0.033	0.010	0.711
E1	0.957	0.405	0.034	0.010	0.714
E2	0.956	0.371	0.041	0.009	0.736
E3	0.968	0.224	0.060	0.006	0.770

(b) Classification tree

Dataset	AUC	FNR	FPR	FOR	FDR
E0	0.927	0.560	0.042	0.013	0.805
E1	0.944	0.172	0.090	0.004	0.824
E2	0.941	0.112	0.113	0.003	0.846
E3	0.970	0.095	0.054	0.002	0.719



(a) MaxEnt



(b) Classification tree

Figure 4: ROC curve for E0 (black solid line), E1 (grey dashed line), E2 (brown dotted line) and E3 (orange longdash line)

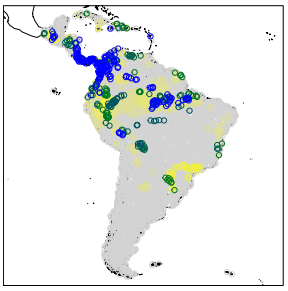


Figure 5: Mapping of the predicted probability of presence rescaled between 0 (grey) and 1 (blue) for the best CT model with tertiles at 1/3 (yellow) and 2/3 (green)

Predictive performance of the models are shown in the discrimination diagnostic plots (Figures 3 and 4). The thresholds in predicted probability of presence optimize total misclassification rate, and occur in the region of 10%. Specifically thresholds were set at—0.11, 0.09, 0.11 and 0.10 for MaxEnt models and 0.145, 0.105, 0.11 and 0.07 for CT models—for the whole dataset and envelopes E1, E2 and E3 respectively.

The error trade-off and frequency distributions (Figure 3a) show that MaxEnt models seem qualitatively unaffected by the amount of 'naughty-noughts' eliminated from the data. The ROC curve (Figure 4a) suggests that the model with greatest control for 'naughty-noughts' performs best, with its curve closest to the top left corner of high sensitivity and high specificity. Interestingly, inspection of the diagnostic performance measures (Table 1a) highlights that eliminating 'naughty-noughts' does improves sensitivity (FNR almost halves) but reduces specificity (FPR almost doubles), and slightly improves AUC. There is small improvement in predictions of presence (FOR nearly halves) and but worse in prediction of absence (FDR increases by 6 points). However, all four MaxEnt models, show poor discrimination as the most common predicted probability is low (below 20%), for *both* absences and presences.

For CT models, we see much more obvious improvements in model performance as more 'naughty-noughts' are eliminated. Overall performance improves, with the ROC (Figure 4b) and AUC indicating clear superiority for fitting the CT SDM within the smallest envelope. Diagnostic performance (Table 1b) improves or remains stable, as sensitivity increases more than five-fold (FNR drops from 56% to 9.5%) and specificity at first increases then marginally increases (FPR changes from 4.2% to 11.3% to 5.4%). Predictive performance improves for both presences (FOR drops sixfold from 1.3% to 0.2%) and absences (FDR drops nine points from 81% to 72%). Moreover, inspection of the error frequency distributions (RHS, Figure 3b) shows that the most common predicted probability of presence for true presences does not overlap with that for true absences.

The best CT model outperforms the best MaxEnt model on all performance measures.

5 CONCLUSIONS

Thus, eliminating 'naughty-noughts' leads to a demonstrated improvement in SDM performance, with a greater improvement and better final performance when a classification tree model is used, rather than MaxEnt within envelopes. This in turn suggests that predictions, for the whole study region, may be biased by 'naughty-noughts'.

We note that simply assessing model performance by inspecting ROC plots and AUC would not discriminate between CT and MaxEnt, and can also obscure poor model performance on other criteria. Indeed, ignoring the fine detail of diagnostic and predictive performance would lead to deducing that the MaxEnt model out-

performs CART, and would also lead to less useful model predictions. In particular sensitivity as well as predictive performance (according to both FOR and FDR) are far superior for the best CT that eliminates most 'naughty-noughts'.

Finally, we note that although there was insufficient room to provide details, the classification tree provides an intuitive model within the environmental envelope that is easily interpretable as it defines environmental profiles corresponding to *Bradypus variegatus* presence or absence. In contrast, MaxEnt provides averaged modelled effects of the (marginal) species response to each environmental gradient independently, which is more difficult to interpret in terms of profiles or habitats.

REFERENCES

Austin, M. and J. Meyers (1996). Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. *Forest Ecology and Management 85*(13), 95 – 106.

Elith, J., M. Kearney, and S. Phillips (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution 1*(4), 330–342.

Elith, J., A. Lehmann, J. Li, et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography 29*(2), 129–151.

Falk, M. G., R. O'Leary, M. Nayak, P. Collins, and S. Low Choy (2014). A Bayesian hurdle model for analysis of an insect resistance monitoring database. *Environmental and Ecological Statistics*, 1–20.

Fielding, A. H. and J. F. Bell (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation 24*(01), 38–49.

Guillera-Arroita, G., J. J. Lahoz-Monfort, and J. Elith (2014). Maxent is not a presence-absence method: a comment on thibaud etal. *Methods in Ecology and Evolution 5*(11), 1192–1197.

Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J. M. C. Overton, R. Aspinall, and T. Hastie (2006). Making better biogeographical predictions of species distributions. *Journal of Applied Ecology 43*(3), 386–392.

Guisan, A., R. Tingley, J. B. Baumgartner, et al. (2013). Predicting species distributions for conservation decisions. *Ecology Letters 16*(12), 1424–1435.

Hijmans, R. J. and J. Elith (2015). Species distribution modeling with r. Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.9177&rep=rep1&type=pdf`.

Low Choy, S. J. (2001). *Hierarchical models for 2D presence/absence data having ambiguous zeroes: With a biogeographical case study on dingo behaviour*. Ph. D. thesis, Queensland University of Technology.

Martin, T. G., P. M. Kuhnert, K. Mengersen, and H. P. Possingham (2005). The power of expert opinion in ecological models using Bayesian methods: impact of grazing on birds. *Ecological Applications 15*(1), 266–280.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine 8*(4), 283 – 298.

Murphy, A. H. and R. L. Winkler (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 26*(1), pp. 41–47.

Oksanen, J. and P. R. Minchin (2002). Continuum theory revisited: what shape are species responses along ecological gradients? *Ecological Modelling 157*(2), 119 – 129.

O'Leary, R. A. (2008). *Informed statistical modelling of habitat suitability for rare and threatened species*. Ph. D. thesis, Queensland University of Technology.

Osborne, P. E. and S. Suárez Seoane (2002). Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling 157*(2-3), 249 – 259.

Pearce, J. and S. Ferrier (2000). An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling 128*(2), 127–147.

Phillips, S. J., R. P. Anderson, and R. E. Schapire (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling 190*(3), 231–259.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications 19*(1), 181–197.

Pirathiban, R., K. J. Williams, J. Murray, and S. J. Low Choy (2015). Eliminating "naughty noughts" in species distribution modelling; does it matter? in preparation for submission to *Environmental Modelling and Software*.

Therneau, T. M., E. J. Atkinson, et al. (1997). An introduction to recursive partitioning using the RPART routines. Available at `http://r.789695.n4.nabble.com/attachment/3209029/0/zed.pdf`.

Williams, K. J. (1998). *Predicting eucalypt distributions in Tasmania: an application of generalised linear modelling*. Ph. D. thesis, University of Tasmania.